JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

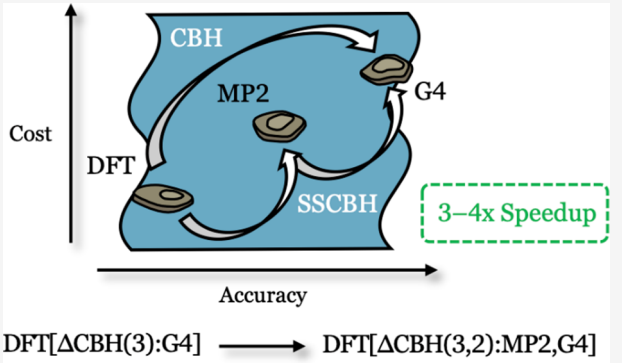
Stepping-Stone CBH: Benchmark and Application of a Multilayered Isodesmic-Based Correction Scheme

Eric M. Collins and Krishnan Raghavachari*

Cite This: J. Chem. Theory Comput. 2024, 20, 3543−3550 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present a generalization of the connectivity-based hierarchy (CBH) of isodesmic-based correction schemes to a multilayered fragmentation platform for overall cost reduction while retaining high accuracy. The newly developed multilayered CBH approach, called stepping-stone CBH (SSCBH), is benchmarked on a diverse set of 959 medium-sized organic molecules. Applying SSCBH corrections to the PBEh-D3 density functional resulted in an average error of 0.76 kcal/mol for the full test set compared to accurate CCSD(T)-quality enthalpies and an even lower error of 0.44 kcal/mol on a subset containing only acyclic molecules. These results rival the traditional CBH-3 approach at a greatly reduced cost, allowing larger fragment corrections to be made at the MP2 level of theory rather than with G4. Our SSCBH approach will enable more widespread applications of CBH methods to a broader range of organic and biomolecular systems.



1. INTRODUCTION

Advancements in computational chemistry now allow the thermodynamic properties of small molecules to be studied at or beyond chemical accuracy (±1 kcal/mol).[1−15] Currently, the well-established method for computing energies with such accuracy is coupled cluster (CC) theory, treating single (S), double (D), triple (T),... excited configurations to accurately capture electron correlation effects.[16] However, combinations of electrons quickly lead to steep computational scaling (ranging from $O(N^6)$−$O(N^{10})$ for CCSD−CCSDTQ), restricting the best of these methods (CCSDTQ5+) to very small molecules or very small basis sets or both.[5,6,13,17−20] An alternative approach involves CCSD(T)/CBS, wherein energies at the "gold standard" CCSD(T) level (scaling as $O(N^7)$) are evaluated at the complete basis set limit and is applicable for small- to medium-sized molecules while retaining chemical accuracy.[16] Unfortunately, even these methods are not directly applicable to many chemically relevant larger molecules.

Effective solutions to this problem come in the form of composite and hybrid (fragmentation) methods. Popular composite wave function-based methods (cWFTs), including Wn,[6,15,21,22] G4,[23] CBS-*n*,[24,25] HEAT,[14] and ccCA,[26] approximate CCSD(T+)/CBS through a series of calculations at various levels of theory and basis set sizes. Although cWFTs have increased the applicability of CCSD(T)/CBS on medium-sized systems, the steep computational scaling of these methods still prohibits universal application. Thus, for large molecules, approximate density functional methods are used almost universally, scaling more moderately, albeit at a significant loss of accuracy.

To mitigate this decrease in accuracy, fragmentation-based methods provide a correction to a low level of theory by first deconstructing the large molecule into smaller overlapping fragments on which more accurate levels of theory can be used readily and then adding fragment energy differences between the two levels of theory to approximate the full system at the higher level of theory. Of particular interest for this work is the Connectivity Based Hierarchy (CBH) of error cancellation schemes. CBH is a hybrid method based on an extension of the isodesmic bond separation scheme.[27] It is also related to ideas developed by other authors on a range of more sophisticated homodesmotic error correction schemes.[28,29] The central idea in such methods is to set up a reaction scheme wherein both reactants and products feature a similar chemical environment, allowing low-level methods to calculate this reaction energy quite accurately due to the cancellation of systematic errors.

CBH is a hierarchy arranged in rungs where each successive rung increasingly preserves more of the chemical environment. The rungs of CBH break down a parent molecule into a set of

Received: December 4, 2023
Revised: March 31, 2024
Accepted: April 1, 2024
Published: April 17, 2024

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

fragments at a given rung n, maintaining the local environment to a specified level based on the local connectivity. As CBH is a generalization of isodesmic reactions, these fragments make up the product side of the CBH reaction. The reactant side is then composed of the original molecule along with overlapping regions of the adjacent product fragments to take double counting into account.

To utilize these schemes for chemical accuracy, all fragments are calculated at a low level of theory as well as at a higher level of theory, and then, the difference between the two is added to the energy of the parent molecule calculated at the low level of theory to approximate the high level, full molecule calculation. This procedure achieves a target mean-absolute error (MAE) of <2 kcal/mol compared to accurate calculations or experimental values for a diverse group of density functionals and <1 kcal/mol for wave function-based methods at as low as the CBH-2 rung.[27,30,31] The success of these CBH corrections on such a wide range of methods lends strong support to the idea of significant and systematic error cancellation starting at the second rung of CBH, which preserves the immediate connectivity of each heavy atom. Lower rungs, including the original isodesmic scheme (CBH-1), typically do not provide sufficient error cancellation, as the associated errors are quite high, >3 kcal/mol, for many systems.

Although average errors of 1−2 kcal/mol have been achieved through CBH-2, better performance can be attained at higher rungs (CBH-3 and CBH-4).[27,30,31] For example, CBH-3 provided a modest improvement in the calculated thermochemical properties of medium-sized biofuel molecules.[32] However, systematic benchmark studies with CBH have not used these higher rungs, as these introduce larger computational costs due to the increase in fragment size, which must be calculated at the high (target) level of theory. The utility of these higher rungs is explored herein using a multistep CBH approach in which larger fragments (CBH-3) are used to construct a correction from a low- to an intermediate-level of theory, and then smaller fragments (CBH-1 or CBH-2) are used to further correct the energy from the intermediate- to a high-level of theory.

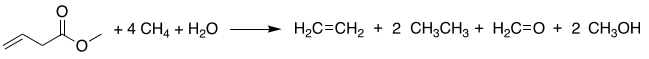
Adding multiple levels of corrections has two potential benefits: (1) reducing the overall computational cost of fragment calculations and (2) increasing the overall accuracy at a similar computational cost. Depending on how well the intermediate level of theory captures long-range effects, the accuracy obtained through the inclusion of an additional step between low- and high-levels of theory will approach the equivalent two-layer model in which the high level is used directly in combination with a higher rung of CBH, but at a significantly reduced cost. Additionally, CBH is limited by the accuracy of its high-level fragment calculations, typically G4 or CCSD(T). The accuracy of G4 is around 1 kcal/mol on average, thus extrapolated energies E(DFT:G4) are designed to achieve a slightly lower accuracy of 1−2 kcal/mol compared to the *true* energy.

This work introduces and benchmarks strategies involving a multilayered CBH approach called Stepping-Stone Connectivity-Based Hierarchy (SSCBH). Just as stepping stones provide a pathway over streams of water, SSCBH utilizes a sequence of smaller steps to reach a target rather than one large step. SSCBH can also be viewed as a sum of interactions broken down into long-, medium-, and short-range, each calculated at an increasingly more accurate level of theory. In such an approach, the highest level of theory only captures local short-range effects, and lower levels of theory must be chosen carefully as they will be responsible for modeling all other longer-range interactions.

The work presented herein is broken into two sections. First, a graph-theoretic analysis of CBH is performed to facilitate an automated set of Python scripts to derive the reactants and products at each order of CBH. A benchmark study on SSCBH is then presented by including intermediate levels of theory, MP2 or CCSD(T) with smaller basis sets, to extrapolate from DFT (low level) to G4 (high level). The calibrations are first carried out on a test set of 28 molecules, followed by a more comprehensive assessment on a much larger test set of 959 molecules.

2. METHODS

2.1. Connectivity-Based Hierarchy. In theoretical thermochemistry, the energies of reactions describing small structural changes can be accurately calculated without the need for highly sophisticated methods due to matching chemical environments between the products and reactants. This phenomenon is due to a high degree of error-cancellation of approximate methods when the chemical environment is conserved between products and reactants. In the 1970s, Pople et al. first popularized these ideas with the introduction of the (isodesmic) bond separation reaction, in which a molecule is separated into its constituent heavy atom bonds. In a bond separation reaction, e.g., methyl 3-butenoate shown below, the number of bonds of a given formal type is retained throughout the chemical transformation.



Thus, the associated energy change (heat of bond separation) can be calculated accurately with inexpensive levels of theory that have systematic errors; i.e., deviations in the bond energies calculated with a given method are similar for similar types of bonds.

The connectivity-based hierarchy (CBH) of isodesmic-type reaction schemes expands on these ideas by defining a generalized set of chemical environment-preserving reactions that can be constructed in a systematic manner based solely on the connectivity of a molecule. There are three components to a CBH reaction: the molecule of interest (parent molecule), the product-side molecules (primary fragments), which preserve the chemical environments of the original system, and the reactant-side molecules (overlap fragments), which balance any overcounted atoms. Together, the energies of the parent and fragment molecules give the reaction energy E(CBH):

$$E(\mathrm{CBH}) = \sum E_{\mathrm{products}} - \sum E_{\mathrm{reactants}} = \sum E_{\mathrm{frag}}^{\mathrm{primary}} - \sum E_{\mathrm{frag}}^{\mathrm{overlap}} - E^{\mathrm{parent}} \quad (1)$$

Since these reaction energies can be calculated with a reasonable accuracy regardless of which level of theory is used, CBH schemes can be further utilized to approximate properties calculated at expensive levels of theory at a greatly reduced computational cost. A correction term can be constructed using the fragment energies calculated at two levels of theory: an approximate (Low) and a sophisticated (High) level of theory.

$$\Delta E = E_{\mathrm{frag}}(\mathrm{High}) - E_{\mathrm{frag}}(\mathrm{Low}) \quad (2)$$

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

$$\Delta\mathrm{CBH(Low: High)} = \sum \Delta E^{\mathrm{primary}} - \sum \Delta E^{\mathrm{overlap}} \quad (3)$$

Under the assumption that $E(\mathrm{CBH})(\mathrm{Low}) \approx E(\mathrm{CBH})(\mathrm{High})$ and a rearrangement of eqs 1−3, the parent molecule energy $E(\mathrm{High})$ can be approximated from a series of significantly less expensive calculations, bypassing the need for the full molecule calculation at the high level of theory.

$$E^{\mathrm{parent}}(\mathrm{High}) \approx E_{\mathrm{CBH}-n}(\mathrm{Low: High}) = E^{\mathrm{parent}}(\mathrm{Low}) + \Delta\mathrm{CBH} - n(\mathrm{Low: High}) \quad (4)$$

Here, the different reaction schemes are denoted by CBH-n, where larger values of n feature a larger amount of preservation. The energy correction is represented as ΔCBH-n(Low: High) and the corrected total energy of the parent molecule as $E_{\mathrm{CBH}\text{-}n}$(Low:High). The quality of the approximated values depends on the chosen preservation scheme of CBH, with larger fragments leading to smaller deviations from the true energy. Typically, deviations within chemical accuracy (defined as 1 kcal mol^{-1}) are achieved with the CBH-2 scheme and above, corresponding to the hypohomodesmotic scheme.

In the current work, we generalize eq 4 to account for more than one correction added. An approximated energy expression utilizing a multistep correction is shown in eq 5 with two CBH corrections from rungs, m and n, where $m \neq n$.

$$E_{\mathrm{SSCBH}(n,m)} = E^{\mathrm{parent}}(\mathrm{Low}) + \Delta\mathrm{CBH} - n(\mathrm{Low: Med}) + \Delta\mathrm{CBH} - m(\mathrm{Med: High}) \quad (5)$$

In eq 5, three different levels of theory are used, denoted as Low, Med, and High, indicating low, medium (intermediate), and high-fidelity levels of theory under the general assumption that the trend in accuracy for the group of methods is High > Med > Low. The resulting energy is represented as $E_{\mathrm{SSCBH(n,\,m)}}$ or more completely as $E_{\mathrm{SSCBH}(n,m)}$[Low:Med:High].

2.2. Implementation of Automated CBH. We have recently developed pyCBH,[33] an open-source package to derive the CBH reactants and products at different CBH rungs in an automated manner (available on Github at https://github.com/colliner/pyCBH). The development of pyCBH was motivated by the inherent systematic structure of the rungs of fragmentation in CBH as well as the need to quickly calculate thousands of CBH corrections in an automated manner. pyCBH employs a graph-theoretic analysis to derive the CBH reactions using an efficient algorithm, and the full details have been outlined elsewhere. Fragments can be formed either from a parent molecule given in Cartesian coordinates or from the SMILES representation for any user-defined CBH rung.

Included with pyCBH is a lookup table of many of the common fragments formed with CBH-0 to CBH-3 along with a database of energies calculated at various levels of theory. If all fragments of a generated CBH reaction are present in the database, then the ΔCBH correction can be computed automatically from the lookup table without the need for further electronic structure calculations.

2.3. Generalization to Multilayered CBH. To utilize two or more CBH corrections, the levels of theory must be ranked according to their performance and how sophisticated the underlying physics is modeled therein. In general, this ranking would normally follow the formal scaling of the methods. Typically, for most molecules, the calculated ordering will be retained, i.e., Low < intermediate < High, but this may not always be the case. DFT has typically been used as the lowest level of theory in previous CBH studies since these methods are among the most widely used approaches for electronic structure calculations. Many popular density functionals adopt a semiempirical approach by selecting a flexible functional form and then fitting the undetermined coefficients to a set of accurate reference values.[34] While DFT is formally exact, most widely used functionals are not systematically improvable; i.e., the addition of more rigorous approximations or constraints does not guarantee an improvement in performance. The accuracies of many wave function-based methods, on the other hand, can usually be improved through a well-defined procedure. For example, the addition of higher-order excitations in the calculation of the energy in coupled cluster (CC) theory is a systematic strategy to approach the true energy. Thus, it is important during the development of SSCBH and other multilayered fragmentation methods for the combination of levels of theory to be chosen carefully and benchmarked on a wide range of molecules, especially when mixing approximate methods, namely, DFT, and accurate wave function-based methods such as coupled cluster theory.

In search of a set of methods that follow the monotonic increase described above, a variety of method combinations were explored. The high level (target) used here is the fourth generation Gaussian-n cWFT method, G4.[23] This composite method is composed of a series of ab initio calculations at different levels of theory, viz., MP2, MP4, and CCSD(T), in combination with various basis sets. Each of these smaller calculations is a potential candidate for acting as the intermediate level of theory since G4 is more accurate than any one of its constituent calculations. These levels of theory, however, could have substantial errors by themselves, due to the use of small basis sets or frozen core approximations. The core−valence correlation term in the G4 protocol, for example, is treated with an all-electron treatment, "MP2(full)" with a large basis set. In fact, the bulk of the correction terms in G4 are derived from MP2 and MP4 with four different basis sets. To avoid discrepancies from these basis set effects, small basis set MP2 and MP4 calculations were not used as the intermediate level of theory. Overall, the MP2(full)/G3LargeXP and CCSD(T)/6-31G(d) portions of G4 were found to be the most suitable intermediate levels of theory for G4-based SSCBH, due to the former employing a large basis set and the latter being responsible for correlation effects beyond MP4. The use of three corrections (four total levels of theory) using both intermediate options is explored below. Additionally, the effect of treating all electrons vs the frozen core approximation is explained in the Supporting Information and Figure S1.

For the low level of theory, some of the most popular density functionals were tested, along with some less expensive semiempirical methods. All low levels of theory used can be categorized into groups based on their computational cost and scaling. Semiempirical methods, including PM6,[35] PM6-D3,[36] and PM7,[37] are the least expensive methods of the bunch. The slightly more expensive group includes the composite X-3c methods HF-3c,[38] PBEh-3c,[39] and B97-3c.[40] These methods employ minimal basis set density functional theory calculations with a set of 3 corrections to account for known deficiencies of the methods. X-3c methods are robust levels of theory to bridge the gap between semiempirical methods and DFT calculations with a large basis set. The remainder of the low-

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

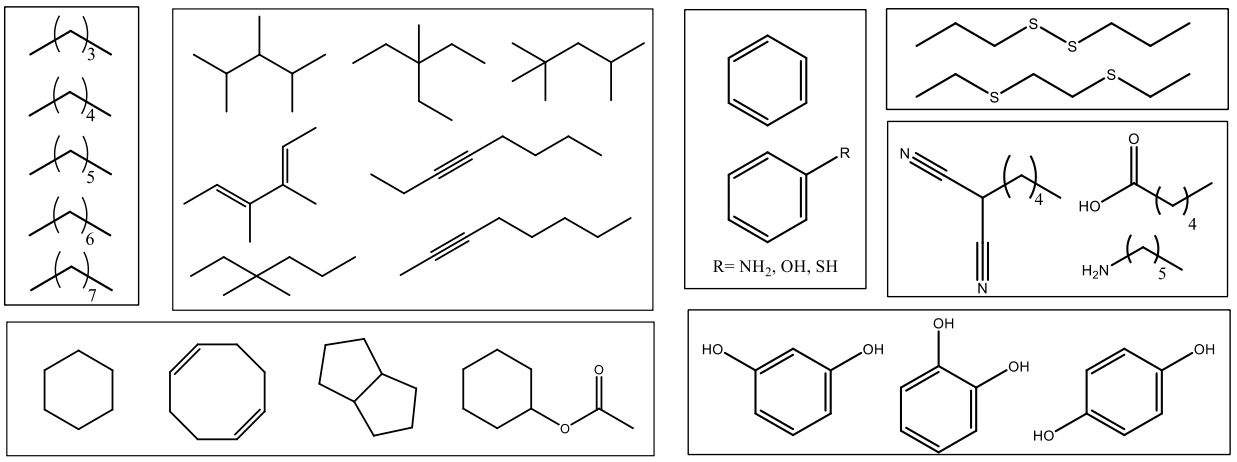


Figure 1. SSCBH Benchmark test set 1 composed of 28 HCNOSCl molecules.

level methods make up the previously most explored group. This group employs standard density functionals paired with the Pople-style basis set 6-311++G(3df,2p).

2.4. Benchmark Data Sets. The first test set (Figure 1) of 28 HCNOSCl-containing organic molecules was used to benchmark which combination of levels of theory works the best for SSCBH. These molecules were chosen as representative of various features in organic molecules, which may be challenging for the standard CBH-2 methodology. Since the largest fragments in this study are from the CBH-3 rung and have a chain length of 4 heavy atoms, parent molecules are required to have a minimum chain length of 5 heavy atoms to avoid capturing the full system in one fragment. Once the new approach was calibrated, the performance of SSCBH was evaluated on a much larger test set of 959 organic molecules. These molecules are a subset of the 1k-G4-C9 data set featuring molecules of up to 9 carbon atoms and up to 13 total heavy atoms with similar compositions as the first test set.[41] All electronic structure calculations in this work were performed with either the Gaussian 16 package[42] or ORCA 4.[43] All of the CBH reactions and corresponding corrections featured in this work were generated with the automated CBH package pyCBH (https://github.com/colliner/pyCBH).

3. RESULTS AND DISCUSSION

3.1. SSCBH Benchmark. The standard CBH protocol (CBH-1, CBH-2, and CBH-3) was applied to the 28-molecule test set as a baseline for benchmarking the multilayered SSCBH (Figure 2). As anticipated, higher rungs of CBH (with increasingly larger fragments used in the correction) decreased the error for all low-level methods, with the performance in the expected order: CBH-3 > CBH-2 > CBH-1. The best semiempirical method for the traditional CBH-3 correction was PM6-D3 with a mean absolute error (MAE) of 4.35 kcal/mol compared to G4 reference values. Minimal basis set composite methods gave impressive results after the CBH-corrections but were slightly worse than the DFT group though at a reduced cost. For comparison, CBH-3-corrected B97-3c and PBE-D3 featured MAEs of 0.98 and 0.68 kcal/mol, respectively. Among the other functionals tested, B3LYP-D3 and ωB97X-D also yielded impressive MAEs with chemical accuracy (less than 1 kcal/mol).

Although the performance is sufficient for many of the low levels of theory after the CBH-3 correction, the required

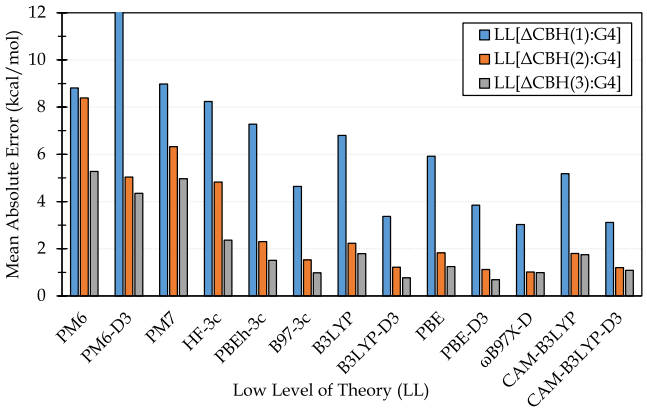


Figure 2. Standard CBH(1−3) protocol for organic molecules of test set 1 using various semiempirical, composite (X-3c), and DFT methods.

calculations can be time-consuming since the energy of these fragments must be calculated at the highest level of theory (G4). SSCBH can reduce the need for these expensive calculations on large fragments by treating such fragments with a lower fidelity method and subsequently correcting them with G4 calculations on smaller fragments. Thus, for the SSCBH approaches illustrated in this work, the highest level of theory (G4) was restricted to either ΔCBH-2 or ΔCBH-1 moving forward. Additionally, we analyzed the importance of dispersion-corrected DFT to capture long-range effects from weak interactions.

First, we consider the performance of methods in which the highest level of theory (i.e., G4) was restricted to ΔCBH-2. Both MP2(Full)/G3LargeXP and CCSD(T)/6-31G(d) were used as the intermediate levels of theory, showing a modest increase in performance for most low levels (LL) of theory with the performance using MP2(full)/G3LargeXP being significantly better than using CCSD(T)/6-31G(d), 0.57 kcal/mol vs 1.59 kcal/mol for PBE-D3. Since CCSD(T) has a much steeper formal scaling than MP2 (N^7 vs N^5), this difference is mostly likely due to basis set effects since the CCSD(T) term in G4 utilizes a much smaller basis set than the MP2(full) term. Therefore, the three-layer models discussed below will mainly focus on MP2(full) as the intermediate level.

The performances of three combinations of multilayered CBH are compared in Figure 3. The best-performing

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

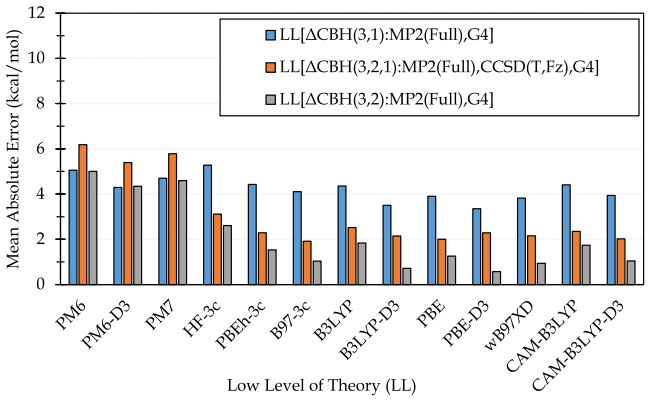


Figure 3. Effect of the intermediate level of theory for multilayered CBH.

combinations, CBH(3,2) starting from B97-3c or most larger basis set DFT methods, achieved MAEs below 1 kcal/mol compared to G4 and are competitive with the corresponding two-layer CBH model in which the high-level fragment calculations are applied with CBH-3. The additional correction from MP2(full) accounts for effects modeled with G4 at a greatly reduced computational cost.

Now, we consider the performance of restricting the highest level of theory (G4) to ΔCBH-1—a three-layer model using MP2(full)/G3LargeXP as the intermediate level of theory as well as a four-layer model using MP2(full)/G3LargeXP for ΔCBH-3 corrections and CCSD(T)/6-31G(d) for ΔCBH-2 corrections. However, restricting the highest level of G4 correction to CBH-1 (isodesmic) proves to be a significant limitation. For example, the CBH(3,1) correction for B3LYP-D3 resulted in a mean absolute error of 3.5 kcal/mol, compared to 0.71 kcal/mol for the CBH(3,2) correction. Although adding additional layers decreases the original CBH-1(LL:G4) errors, the error cancellation from adding CBH-3 and CBH-2 corrections is insufficient to achieve chemical accuracy. Similar discrepancies have been pointed out in previous CBH studies, where the CBH-1 fragments are inadequate to preserve enough of the local chemical environment. The deviation of CBH(3,1) can be reduced with the inclusion of a fourth layer correction, CCSD(T) with CBH-2; however, the errors coming from the small CBH-1 fragments are now combined with errors coming from small basis sets leading to only a slightly better model with a MAE of 2.14 kcal/mol for B3LYP-D3. Finally, a different four layer model using CCSD(T)/6-31G(d) for ΔCBH-3 corrections and MP2(full)/G3LargeXP for ΔCBH-2 corrections (i.e., switching CCSD(T)/6-31G(d) and MP2(full)/G3LargeXP), was also considered (not shown). While the results improved slightly relative to the original four-layer model for non-benzenoid systems, they were significantly worse for benzenoid systems, making the overall performance to be worse. These results appear to be related to the relative signs of the contributions from the two intermediate corrections. If they have opposite signs, there is some cancellation and the composite results are better. If they have the same signs, the composite results get worse. Overall, the best combinations of these methods reach around 2 kcal/mol compared to G4, indicating that the high-level G4 correction must be at a higher rung than CBH-1.

The inclusion of long-range interactions, i.e., dispersion, in the low level of theory is shown to be important since these effects are not corrected for in any of the smaller fragment calculations. The most common dispersion corrections for DFT including Grimme's D3 and D4,[44] along with NL (VV10),[45] all work well. Standard CBH as well as SSCBH are compared to their nondispersion corrected counterparts in Figure 4. In every case, the dispersion-corrected functional

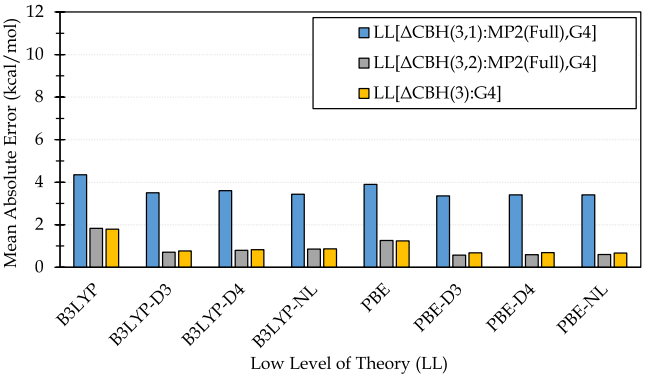


Figure 4. Effect of the various dispersion corrections on the low level of theory for multilayered CBH.

outperformed its standard counterpart for the three-layer models. The top performing combinations of SSCBH achieve between 0.50 and 0.75 kcal/mol average errors with explicit dispersion corrected methods compared to 1−2 kcal/mol errors for standard DFT, highlighting the importance of capturing such long-range effects within the low level of theory. Interestingly, there is no significant difference in the performance between the different dispersion models.

Our overall conclusion from the SSCBH benchmark set is that the best-performing model is to use G4 to get the ΔCBH-2 correction and to use MP2(full)/G3LargeXP as the intermediate level of theory to obtain the ΔCBH-3 correction. This three-layer model, denoted as SSCBH(3,2)[LL:MP2-(full):G4] will now be compared to the ΔCBH-3 model directly obtained with G4 for a much larger test set to assess the overall performance and outlook.

3.2. Application to the 1k-G4-C9 Data Set. The results summarized in Table 1 show the overall mean absolute error of SSCBH corrected DFT and traditional 2-layer CBH. Similar trends are seen on the larger test set of 959 organic molecules. In general, the SSCBH model SSCBH(3,2)[LL:MP2(full):G4] reached a similar performance (within 0.3 kcal/mol) to its two-layer counterpart CBH-3[LL:G4] at a reduced computational cost. There is a significant improvement compared to the two-

Table 1. Mean Absolute Errors (kcal/mol) for Three-layer SSCBH Methods Compared to the Standard Two-Layer CBH Protocol

rung of correction	2 layer	3 layer	2 layer
CBH-1			
CBH-2		G4	G4
CBH-3	G4	MP2(full)	
low level (LL)		kcal mol^{-1}	
PBEh	1.04	1.30	1.61
PBEh-D3	0.65	0.76	1.38
ωB97X-D	0.86	1.14	1.37
CAM-B3LYP	1.57	1.87	1.99
CAM-B3LYP-D3(BJ)	0.83	1.11	1.51

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

layer CBH-2 model, indicating the MP2 correction with CBH-3 is adding additional error cancellation, but not as much as the full CBH-3 G4 correction. Thus, these multilayered corrections can be used to reduce the computational cost of the CBH-3 model with a small loss in accuracy.

These improvements are generalized over a variety of functional groups and molecular features, summarized in Table 2. The multilayer CBH(3,2)[ωB97X-D:MP2(full):G4] model performs with an MAE of 1.14 kcal/mol compared to 0.86 kcal/mol for the CBH-3 model. Overall, these models perform best on the acyclic molecules and hydrocarbons as both subgroups have a significantly lower deviation than the full data set. The performance of the multilayer CBH(3,2)[ωB97X-D:MP2(full):G4] on acyclic molecules is closer to the corresponding performance of the CBH-3[ωB97X-D:G4] as both have similar performances (0.35 vs 0.33), implying that the multilayer model can fully approximate the CBH-3 model at a reduced cost. However, this effect is not seen for conjugated and heterocyclic systems, which feature larger errors regardless of how many layers of CBH corrections are added. Local connectivity-based corrections oftentimes perform worse on these systems due to the nature of their valence bond structures and delocalization effects, which are difficult to capture within the corresponding fragments. Additionally, each fragment as defined by CBH is optimized to the lowest energy conformer, which can introduce conformational differences between the optimized structure of the full molecule and its corresponding fragments, especially with larger fragments at higher rungs of CBH.

Table 2. Mean Absolute Errors (kcal/mol) for SSCBH and CBH Methods for Various Functional Groups and Molecular Features for ωB97X-D

rung of correction	2 layer	3 layer
CBH-1		
CBH-2		G4
CBH-3	G4	MP2(full)
total MAE	0.86	1.14
acyclic	0.35	0.33
alicyclic	1.39	2.12
conjugated	2.10	3.18
heterocyclic	1.34	1.62
hydrocarbons	0.57	0.71
O-containing	0.85	1.15
N-containing	1.26	1.59
S-containing	0.75	0.95
Cl-containing	1.15	1.90

The explicit treatment of dispersion in the low level of theory diminishes the errors to some extent, as the error for ring-containing molecules is slightly above 1.0 kcal/mol for dispersion-corrected PBEh and around 2.5 kcal/mol for the uncorrected counterparts (Tables 3 and 4). The results further reinforce the importance of sufficiently modeling noncovalent interactions in the full molecule calculations at the low level of theory. Dispersion-corrected low-level PBEh-D3(BJ) performed well regardless of which heteroatoms were present in the molecule, although the dispersion correction improved the errors of molecules containing chlorine the most out of any heteroatom, reducing the MAE from 1.93 to 0.86 kcal/mol.

While SSCBH works well in most cases, there are some cases where the single and multilayered CBH schemes break.

Table 3. Mean Absolute Errors (kcal/mol) for SSCBH and CBH Methods for Various Functional Groups and Molecular Features for PBEh

rung of correction	2 layer	3 layer
CBH-1		
CBH-2		G4
CBH-3	G4	MP2(full)
total MAE	1.04	1.30
acyclic	0.69	0.64
alicyclic	1.78	2.64
conjugated	1.61	2.64
heterocyclic	2.07	2.42
hydrocarbons	1.32	1.62
O-containing	1.30	1.58
N-containing	1.61	1.96
S-containing	0.94	1.10
Cl-containing	1.17	1.93

Table 4. Mean Absolute Errors (kcal/mol) for SSCBH and CBH Methods for Various Functional Groups and Molecular Features for PBEh-D3(BJ)

rung of correction	2 layer	3 layer
CBH-1		
CBH-2		G4
CBH-3	G4	MP2(full)
total MAE	0.65	0.76
acyclic	0.44	0.44
alicyclic	0.78	1.04
conjugated	0.94	1.16
heterocyclic	0.97	1.09
hydrocarbons	0.50	0.58
O-containing	0.66	0.81
N-containing	0.87	0.99
S-containing	0.97	0.60
Cl-containing	0.35	0.86

Examples involve (1) larger strained and cyclic molecules and (2) error-cancellation mismatch between the CBH rungs.

1. For some heterocyclic aromatic systems (e.g., 2-aminopyridine), CBH-3 is less reliable without special care, i.e., CBH-2 correction schemes result in smaller errors than CBH-3. We hypothesize this effect is from the optimized geometry of CBH-2 fragments more closely matching the geometries in the parent molecule than that of the CBH-3 fragments. To address this, the parent geometry must be retained as closely as possible for CBH-3 fragments. A possible solution is to consider multiple conformers and the fragment conformer that matches the parent geometry the closest could be used.
2. Another instance in which the SSCBH scheme breaks down is in the case of a mismatch between errors in CBH rungs. One class of molecules in which this is apparent is multisubstituted chlorobenzene (e.g., pentachlorobenzene). The overall results are related to the signs of the intermediate corrections from the two levels. If they have opposite signs, there is some cancellation, and the composite results are better. Overall, the combination of levels of theory used is of the utmost importance in order to avoid mismatching errors and other effects that may appear in certain CBH rungs. Furthermore, using a combination of DFT and wave

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550

function-based methods could also be responsible for an increased chance of observing this phenomenon.

Finally, to demonstrate the efficiency of the new models proposed in this study we selected two 15 heavy-atom test molecules, one aromatic (formula $C_9N_5SH_{17}$), and the other aliphatic (formula $C_9O_6H_{14}$). In both systems, the fragment calculations involved in the three-layer SSCBH(3,2) models (DFT:MP2(full):G4) are 3−4 times faster than the fragment calculations in the two-layer CBH-3 models (DFT:G4). This is primarily because larger CBH-3 fragments (up to 8 heavy-atoms) are considered at the G4 level in the two-layer calculations compared to the smaller CBH-2 fragments (only up to 5 heavy atoms) in the three-layer calculations. These speedups are in addition to the more than 1 order of magnitude speedup of the two-layer CBH-3 calculations relative to the direct G4 calculations.

4. CONCLUSIONS

The benchmarks presented here demonstrate that the multilayered CBH model SSCBH could be used to reduce the computational scaling of higher rungs of CBH corrections, replacing many of the expensive calculations on large fragments with the lower MP2(full) level of theory. Cheaper methods, such as PM6-D3 and X-3c, have been used as the low level of theory with moderate errors compared to G4, indicating that larger molecules could be studied using the SSCBH approach in which DFT is not a viable option. Overall, the accuracy of G4 can be rivaled with a wide variety of less expensive methods at a greatly reduced cost with the SSCBH protocol, with improved accuracy compared to the traditional CBH correction scheme. Additionally, corrections from SSCBH can be utilized to reduce the computational cost of more accurate composite methods. With these methods, the study of larger molecules could be performed without the need for expensive calculations on large fragments. Using a suitable combination of methods, important long-range effects can be captured at intermediate levels of theory with little to no loss in accuracy compared to the more expensive standard CBH approach.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01330.

Computational details, effect of frozen core MP2 for fragment calculations, and coordinates of optimized geometries (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Author

Eric M. Collins − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0002-9113-1705

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01330

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge support from the National Science Foundation grants CHE-1665427 and CHE-2102583 at Indiana University. Big Red II supercomputing facility at Indiana University was used for most of the calculations in this study.

■ REFERENCES

(1) Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *WIREs Computational Molecular Science* **2016**, *6*, 292−310.

(2) Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: Current strategies and future challenges. *Theor. Chem. Acc.* **2012**, *131*, 1079.

(3) Dixon, D. A.; Feller, D.; Peterson, K. A. Chapter one - a practical guide to reliable first principles computational thermochemistry predictions across the periodic table. In *Annual reports in computational chemistry*, Wheeler, R. A., Ed.; Elsevier, 2012; Vol. *8*, pp 1−28.

(4) Helgaker, T.; Klopper, W.; Tew, D. P. Quantitative quantum chemistry. *Mol. Phys.* **2008**, *106*, 2107−2143.

(5) Karton, A.; Daon, S.; Martin, J. M. L. W4−11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles w4 data. *Chem. Phys. Lett.* **2011**, *510*, 165−178.

(6) Karton, A.; Sylvetsky, N.; Martin, J. M. L. W4−17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *J. Comput. Chem.* **2017**, *38*, 2063−2075.

(7) Karton, A.; Tarnopolsky, A.; Martin, J. M. L. Atomization energies of the carbon clusters c n (n = 2−10) revisited by means of w4 theory as well as density functional, gn, and cbs methods. *Mol. Phys.* **2009**, *107*, 977−990.

(8) Klopper, W.; Ruscic, B.; Tew, D. P.; Bischoff, F. A.; Wolfsegger, S. Atomization energies from coupled-cluster calculations augmented with explicitly-correlated perturbation theory. *Chem. Phys.* **2009**, *356*, 14−24.

(9) Karton, A.; Gruzman, D.; Martin, J. M. L. Benchmark thermochemistry of the cnh2n+2 alkane isomers (n = 2−8) and performance of DFT and composite ab initio methods for dispersion-driven isomeric equilibria. *J. Phys. Chem. A* **2009**, *113*, 8434−8447.

(10) Karton, A.; Chan, B. Accurate heats of formation for polycyclic aromatic hydrocarbons: A high-level ab initio perspective. *Journal of Chemical & Engineering Data* **2021**, *66*, 3453−3462.

(11) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gn theory. *WIREs Comput. Mol. Sci.* **2011**, *1*, 810−825.

(12) Simmie, J. M.; Sheahan, J. N. Validation of a database of formation enthalpies and of mid-level model chemistries. *J. Phys. Chem. A* **2016**, *120*, 7370−7384.

(13) Chan, B.; Radom, L. W3x: A cost-effective post-ccsd(t) composite procedure. *J. Chem. Theory Comput* **2013**, *9*, 4769−4778.

(14) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. Heat: High accuracy extrapolated ab initio thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599−11613.

(15) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 theory for computational thermochemistry: In pursuit of confident sub-kj/mol predictions. *J. Chem. Phys.* **2006**, *125*, No. 144108.

(16) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.

(17) Feller, D.; Peterson, K. A.; Dixon, D. A. A survey of factors contributing to accurate theoretical predictions of atomization energies and molecular structures. *J. Chem. Phys.* **2008**, *129*, No. 204105.

(18) Karton, A.; Kaminker, I.; Martin, J. M. L. Economical post-ccsd(t) computational thermochemistry protocol and applications to some aromatic compounds. *J. Phys. Chem. A* **2009**, *113*, 7610−7620.
(19) Karton, A.; Taylor, P. R.; Martin, J. M. L. Basis set convergence of post-ccsd contributions to molecular atomization energies. *J. Chem. Phys.* **2007**, *127*, No. 064104.
(20) Karton, A. Effective basis set extrapolations for ccsdt, ccsdt(q), and ccsdtq correlation energies. *J. Chem. Phys.* **2020**, *153*, No. 024102.
(21) Martin, J. M. L.; de Oliveira, G. Towards standard methods for benchmark quality ab initio thermochemistry—w1 and w2 theory. *J. Chem. Phys.* **1999**, *111*, 1843−1856.
(22) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. W3 theory: Robust computational thermochemistry in the kj/mol accuracy range. *J. Chem. Phys.* **2004**, *120*, 4129−4141.
(23) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, No. 084108, DOI: 10.1063/1.2436888.
(24) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A complete basis set model chemistry. Vii. Use of the minimum population localization method. *J. Chem. Phys.* **2000**, *112*, 6532−6542.
(25) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. A complete basis set model chemistry. V. Extensions to six or more heavy atoms. *J. Chem. Phys.* **1996**, *104*, 2598−2619.
(26) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. The correlation consistent composite approach (ccca): An alternative to the Gaussian-n methods. *J. Chem. Phys.* **2006**, *124*, No. 114104.
(27) Ramabhadran, R. O.; Raghavachari, K. Theoretical thermochemistry for organic molecules: Development of the generalized connectivity-based hierarchy. *J. Chem. Theory Comput* **2011**, *7*, 2094−2103.
(28) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. Alternative approach to problem of assessing stabilization energies in cyclic conjugated hydrocarbons. *Theor Chim Acta* **1975**, *38*, 121−129.
(29) Wheeler, S. E.; Houk, K. N.; Schleyer, P. V. R.; Allen, W. D. A hierarchy of homodesmotic reactions for thermochemistry. *J. Am. Chem. Soc.* **2009**, *131*, 2547−2560.
(30) Ramabhadran, R. O.; Raghavachari, K. Connectivity-based hierarchy for theoretical thermochemistry: Assessment using wave function-based methods. *J. Phys. Chem. A* **2012**, *116*, 7531−7537.
(31) Chan, B.; Collins, E.; Raghavachari, K. Applications of isodesmic-type reactions for computational thermochemistry. *WIREs Computational Molecular Science* **2021**, *11*, No. e1501.
(32) Debnath, S.; Sengupta, A.; Raghavachari, K. Eliminating systematic errors in DFT via connectivity-based hierarchy: Accurate bond dissociation energies of biodiesel methyl esters. *J. Phys. Chem. A* **2019**, *123*, 3543−3550.
(33) Raghavachari, K.; Maier, S.; Collins, E. M.; Debnath, S.; Sengupta, A. Approaching coupled cluster accuracy with density functional theory using the generalized connectivity-based hierarchy. *J. Chem. Theory Comput* **2023**, *19*, 3763−3778.
(34) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(35) Stewart, J. J. P. Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(36) Řezáč, J.; Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput* **2012**, *8*, 141−151.
(37) Stewart, J. J. P. Optimization of parameters for semiempirical methods vi: More modifications to the nddo approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1−32.
(38) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(39) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, No. 054107.
(40) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97−3c: A revised low-cost variant of the b97-D density functional method. *J. Chem. Phys.* **2018**, *148*, No. 064104.
(41) Collins, E. M.; Raghavachari, K. Effective molecular descriptors for chemical accuracy at DFT cost: Fragmentation, error-cancellation, and machine learning. *J. Chem. Theory Comput* **2020**, *16*, 4938−4950.
(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 rev. C.01*; Gaussian Inc.: Wallingford, CT, 2016.
(43) Neese, F. Software update: The orca program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8*, No. e1327, DOI: 10.1002/wcms.1327.
(44) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements h-pu. *J. Chem. Phys.* **2010**, *132*, No. 154104.
(45) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der waals density functional: The simpler the better. *J. Chem. Phys.* **2010**, *133*, No. 244103.

https://doi.org/10.1021/acs.jctc.3c01330
J. Chem. Theory Comput. 2024, 20, 3543−3550